

Generating Grasp Poses for a High-DOF Gripper Using Neural Networks

Min Liu^{1,4}, Zherong Pan², Kai Xu^{1*}, Kanishka Ganguly³, and Dinesh Manocha⁴

https://gamma.umd.edu/researchdirections/grasping/high_dof_grasping

Abstract—We present a learning-based method for representing grasp poses of a high-DOF hand using neural networks. Due to redundancy in such high-DOF grippers, there exists a large number of equally effective grasp poses for a given target object, making it difficult for the neural network to find consistent grasp poses. We resolve this ambiguity by generating an augmented dataset that covers many possible grasps for each target object and train our neural networks using a consistency loss function to identify a one-to-one mapping from objects to grasp poses. We further enhance the quality of neural-network-predicted grasp poses using a collision loss function to avoid penetrations. We use an object dataset that combines the BigBIRD Database, the KIT Database, the YCB Database, and the Grasp Dataset to show that our method can generate high-DOF grasp poses with higher accuracy than supervised learning baselines. The quality of the grasp poses is on par with the groundtruth poses in the dataset. In addition, our method is robust and can handle noisy object models such as those constructed from multi-view depth images, allowing our method to be implemented on a 25-DOF Shadow Hand hardware platform.

I. INTRODUCTION

Grasp pose generation and prediction are important problems in robotics [35], [34], [18]. Recently, learning-based methods [23], [19], [20] have achieved high rates of success in terms of grasping unknown objects. However, these methods are mainly limited to low-DOF grippers with only 1-6 DOFs or they assume that a high-DOF gripper moves in a low-DOF subspace [7]. This assumption limits the space of the grasp poses a robot hand can represent and the space of the target objects the hand can handle. In this work, we address the problem of developing learning algorithms for generating grasp poses for a high-DOF hand. Such high-DOF hands have been used to perform complex in-hand manipulations in prior works [2], [30].

Generating grasp poses for a high-DOF gripper is more challenging than for low-DOF grippers due to the existence of pose ambiguity, i.e. there exists a large number of equally effective grasp poses for a given target object. However, we need to train a single neural network to predict one grasp pose for each object. As a result, we need to pick a set of grasp poses for a set of target objects, that can be represented by the neural network. In the case of a low-DOF gripper, if the neural network predicts the correct direction

and orientation towards the object, one can simply close the gripper and the predicted grasp operation will very likely be successful. Therefore, most prior works [23], [37], [7] only learn the approaching direction and orientation of the gripper. For a high-DOF gripper, however, there are multiple remaining DOFs (beyond direction and orientation) to be determined after the wrist pose is known. Computing these remaining DOFs is still a major challenge in deep-learning-based grasp pose generation methods.

There are two kinds of learning-based methods for grasp pose generation. In the first [20], [10], a grasp pose is generated using two steps. First, a neural network is trained to predict the possibility of success given a grasp pose as an input. Second, the grasp pose is generated during runtime using a sampling-based optimizer such as a multi-armed bandit [24] to maximize the possibility of success. This method does not suffer from pose ambiguity because it allows multiple grasps to be equally effective for a single object. However, the high-DOF nature of the gripper results in a large search space for the sampling-based optimizer, making the online phase very computationally costly. In the second kind of method [7], a neural network is trained to predict the grasp poses directly from single-view observations of the object. As a result, this direct method becomes very efficient because only a forward propagation through the neural network is needed to generate the grasp pose. However, since many high-DOF grasp poses can be equally effective for a single object, an additional constraint is required to guide neural networks to determine the poses from which it should learn. Due to the lack of such guidance, [7] can not be used to generate high-DOF grasp poses directly.

Main Results: We present a learning-based method for representing grasp poses for a high-DOF articulated robot hand. Our method enables fast grasp pose generation without the low-DOF assumption. Similar to [7], we train a neural network to predict grasp poses directly so that grasp poses can be generated efficiently during runtime. To resolve the ambiguity of grasp poses for each object, we introduce the notion of **consistency loss**, which allows the neural network to choose from a large number of candidate grasp poses and select the one that can be consistently represented by a single neural network. However, the grasp pose predicted by the network can be in close proximity to the object, leading to many hand-object penetrations. To resolve this issue, we introduce collision loss, which penalizes any robot-object penetrations, to push the gripper outside the object. We train the neural network for 40 hours on a dataset of 324 objects by combining the BigBIRD Database [33], the KIT Object Models Database [15], the YCB Benchmarks [6], and the Grasp Database [14]. We show that our method can achieve 4× higher accuracy (in terms of distances to the

¹Min Liu and Kai Xu are with School of Computer, National University of Defense Technology. {gfsliumin@gmail.com, kevin.kai.xu@gmail.com}

²Zherong is with Department of Computer Science, University of North Carolina at Chapel Hill. {zherong@cs.unc.edu}

³Kanishka Ganguly is with UMIACS (Institute for Advanced Computer Studies), University of Maryland at College Park. {kganguly@umiacs.umd.edu}

⁴Min Liu and Dinesh Manocha are with Department of Computer Science and Electrical & Computer Engineering, University of Maryland at College Park. {dm@cs.umd.edu}

*Kai Xu is the corresponding author.

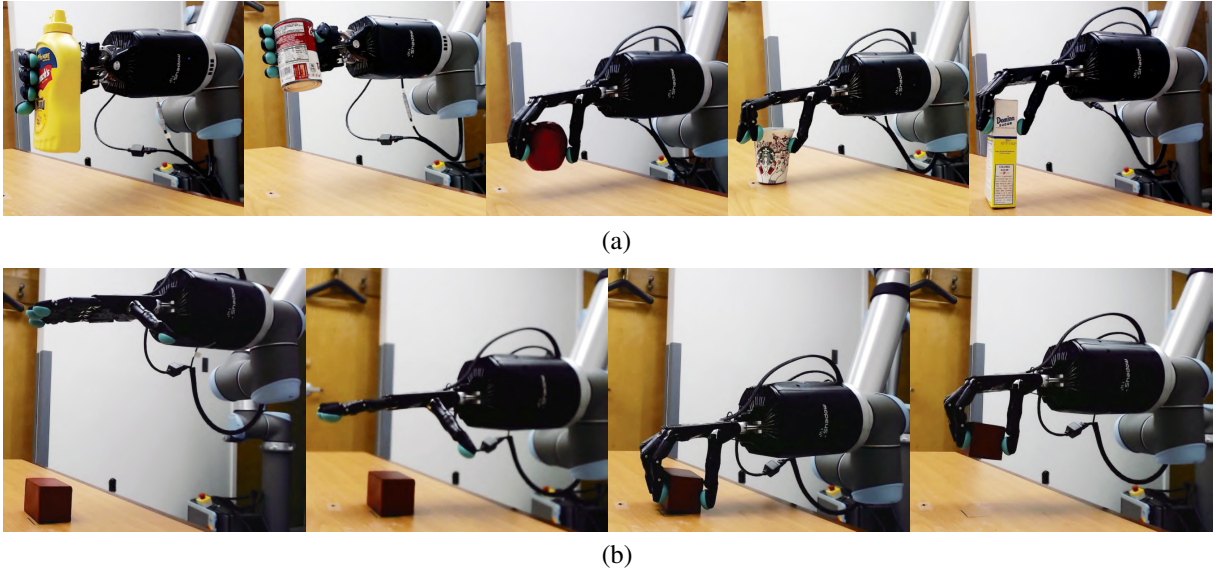


Fig. 1: (a): Our method implemented on the 25-DOF Shadow Hand to grasp different objects. (b): Several frames of a single grasp process.

groundtruth grasp poses) than supervised learning baselines in grasp pose representation. In addition, we show that our method can be used in several application scenarios by taking inaccurate 3D object models as inputs; these object models are reconstructed from multi-view depth images. As a result, our method can be implemented on 25-DOF Shadow Hand hardware, as shown in Figure 1, where each grasp can be computed within 3-5 seconds at runtime.

The rest of the paper is organized as follows. We review related work in Section II and then formulate our problem in Section III. The main neural network architecture and training algorithm are presented in Section IV. Finally, we highlight the performance on different objects in Section V.

II. RELATED WORK

Methods for robot grasp pose generation can be classified based on the assumptions they make about the inputs. Early works [38], [25], [4], [26] are designed for complete 3D shapes, such as 3D triangulated meshes of objects, as inputs. To estimate the quality of a grasp pose [38], [25] or compute a feasible motion plan [4], [26] deterministically, a 3D mesh representation is used. However, these methods are difficult to deploy on current real-world grasping systems or robot hands due to discrepancies and sensing uncertainties. Most practical grasp planning methods that can take incomplete shapes are based on machine learning. Early learning methods predict good grasp poses [12] or points [32] from several RGB images using manually engineered features and supervised/active learning [27]. More recently, learning-based grasp pose prediction algorithms [24], [23], [20], [7], [5], [13], [29] replace manually engineered features with features learned from deep convolutional networks for better generality and robustness. All these methods are designed for low-DOF grippers. Some learning-based methods [1], [36] take an incomplete or a partial shape as input and internally reconstruct a voxelized shape. Our method uses [26] to

generate groundtruth grasping data and we assume that the input to the neural-network is a complete object model represented using an occupancy grid. However, our trained network is robust to data inaccuracies and can be applied to object models reconstructed from multi-view depth images.

Most existing learning-based methods [12], [24], [23], [20], [7], [21] use the learned model in a two-stage algorithm. During the first step, the learned neural network takes both the observation of the object and a proposed grasp pose as inputs and predicts the possibility of a successful grasp. During the second step, the final grasp pose is optimized to maximize the rate of success using exhaustive search [12], gradient-based optimization [21], [22], sampling-based optimization [20], or multi-armed bandits [24]. Instead, our method uses a learning model to predict the grasp poses for a high-DOF gripper directly. Our method is similar to [7], which learns a neural network to predict the grasp poses directly, but [7] is designed for low-DOF grippers and pose ambiguity is not handled.

III. PROBLEM FORMULATION

In this section, we formulate the problem of high-DOF grasp pose generation. Each grasp pose is identified with a high-DOF configuration of the robot hand $\mathbf{x} = (\mathbf{x}_b^T \mathbf{x}_j^T)^T$, where \mathbf{x}_b is the 7-DOF rigid transformation of the hand wrist and \mathbf{x}_j is the remaining DOFs, i.e., joint angles. Our goal is to find a mapping function $f(\mathbf{o}) = \mathbf{x}$, where \mathbf{o} is an observation of the object \mathcal{O} . This observation can take several forms. In this paper, we assume that \mathbf{o} is the 3D occupancy grid [8], [11] derived by discretizing the object. We denote \mathbf{o}_s as the signed distance field [28] derived by solving the Eikonal equation from the original mesh.

We use deep neural networks to represent f with optimizable parameters denoted by θ . The main difference between our method and prior deep-learning-based methods [36], [10], [24] is that our network directly outputs the

grasp pose \mathbf{x} . Prior methods only predict the possibility of successful grasps, given a possible grasp pose, which can be summarized as a function $g(\mathbf{x}, \mathbf{o}) = p$, where p is the possibility of success. Function g has advantages over our function f because g allows multiple versions of \mathbf{x} to be generated for a single \mathcal{O} . However, to use g , we need to solve the following problem:

$$\underset{\mathbf{x}}{\operatorname{argmax}} \quad g(\mathbf{x}, \mathbf{o}), \quad (1)$$

which can be computed efficiently for low-DOF grippers using either sampling-based optimization [20] or multi-armed bandits [24]. However, this optimization can be computationally costly for a high-DOF gripper due to the high-dimensional search space. This optimization can also be ill-posed and under-determined because many unnatural grasp poses might also lead to effective grasp poses, as shown in [9]. This is our main motivation for choosing f over g . However, training a neural network that represents function f is more challenging than training g for two reasons.

- If we have a dataset of N objects and groundtruth grasp poses $\{<\mathcal{O}_i, \mathbf{x}_i>\}$, a simple training method is to use the data loss $\mathcal{L}_{data} = \sum_i \|f(\mathbf{o}_i, \boldsymbol{\theta}) - \mathbf{x}_i\|^2$. However, since multiple grasp poses \mathbf{x}_i are valid for each object \mathcal{O}_i , we can build many datasets for the same set of objects $\{\mathcal{O}_i\}$ by choosing different grasp poses for each object. The resulting data loss \mathcal{L}_{data} generated by using different datasets can be considerably different according to our experiments. Therefore, the first challenge in training function f is that we need to build a dataset leading to a small \mathcal{L}_{data} after training.
- A second problem in training f is that we have to ensure the quality of grasp poses generated by the neural networks. The quality of a grasp pose in learning-based methods can be measured by comparing it with the groundtruth pose. However, there are other important metrics. For example, a grasp pose should not have penetration with \mathcal{O} . In prior methods [36], [10], [24], the neural network is not responsible for ensuring the quality of the grasp poses, but we can guarantee high-quality grasp poses when solving Equation 1 after training. However, in our case, the neural network is used to generate \mathbf{x} directly, so our final results are very sensitive to the outputs of the neural network.

IV. LEARNING HIGH-DOF GRASP POSES

In this section, we present the architecture of our neural network used for high-DOF grasping.

A. Neural Networks

We represent f using a deep neural network, as illustrated in Figure 2. We assume that a high-DOF grasp pose can be generated from a low-dimensional feature vector of the object denoted by $\boldsymbol{\omega}$; a similar approach is used by [7]. We use a fully connected sub-network $\text{NN}_{\mathbf{x}}$ to parameterize this mapping function:

$$\mathbf{x} = \text{NN}_{\mathbf{x}}(\boldsymbol{\omega}, \theta_1),$$

where θ_1 is the optimizable weights. To parameterize $\text{NN}_{\mathbf{x}}$, we use a network with 3 hidden layers with $(64 \times 7 \times 7 \times 7 =) 21952, 4096, 1024$ neurons, respectively. We use ReLU activation functions for each hidden layer and we add batch normalization to the first two hidden layers. When different sensors leading to different observations of \mathcal{O} (e.g., an occupancy grid or a depth image,) are used in our application, we use another sub-network to transform the observation to $\boldsymbol{\omega}$. Therefore, we have:

$$\text{NN}_{\mathbf{o}}(\mathbf{o}, \theta_2) = \boldsymbol{\omega},$$

and $\theta \triangleq (\theta_1^T \theta_2^T)^T$. This neural network is fully convolutional. $\text{NN}_{\mathbf{o}}$ has 3 3D-convolutional layers with 64 kernels of size 4. We add batch normalization, ReLU activation, and max-pooling layers after each convolutional layer. Finally, we have $f = \text{NN}_{\mathbf{x}} \circ \text{NN}_{\mathbf{o}}$.

B. Consistency Loss

In practice, optimizing θ is difficult due to the two challenges discussed in Section III. We resolve these issues using two loss functions. Our first loss function is called a consistency loss function and we use this function used to resolve the grasp pose ambiguity for each \mathcal{O} . Instead of picking one grasp pose \mathbf{x}_i for each \mathcal{O}_i during dataset construction, we compute a set of K grasp poses denoted by $\mathbf{x}_{i,j}$ for each \mathcal{O}_i , where $j = 1, \dots, K$, resulting in a large dataset with NK grasp poses for N objects. As a result, our consistency loss function takes the following form:

$$\mathcal{L}_{consistency} = \sum_i \min_j \|f(\mathbf{o}_i, \boldsymbol{\theta}) - \mathbf{x}_{i,j}\|^2 / N.$$

This novel formulation allows the neural network to pick the N grasp poses leading to the smallest residual. Note that, although $\mathcal{L}_{consistency}$ is not uniformly differentiable, its sub-gradient exists and optimizing $\mathcal{L}_{consistency}$ with respect to both θ and j can be performed with the conventional back-propagation gradient computation framework [16]. Specifically, after forward propagation computes $f(\mathbf{o}_i, \boldsymbol{\theta})$ for every i , we pick j leading to the smallest residual, and finally perform backward propagation with:

$$\frac{\partial \mathcal{L}_{consistency}}{\partial f(\mathbf{o}_i, \boldsymbol{\theta})} = (f(\mathbf{o}_i, \boldsymbol{\theta}) - \mathbf{x}_{i,j^*}) / N$$

$$j^* = \underset{j}{\operatorname{argmin}} \|f(\mathbf{o}_i, \boldsymbol{\theta}) - \mathbf{x}_{i,j}\|^2.$$

C. Collision Loss

To resolve the second challenge and ensure the quality of the learned grasp poses, we note that most incorrect or inaccurate \mathbf{x} predicted by the neural network have the gripper intersecting with \mathcal{O} . To resolve this problem, we add a second loss function that penalizes any penetrations between the gripper and \mathcal{O} . Specifically, we first construct a signed distance function \mathbf{o}_s from the original mesh and then sample a set of points $\mathbf{p}_{i=1, \dots, P}$ on the gripper. Next, we formulate the collision loss function as:

$$\mathcal{L}_{collision} = \sum_{i=1}^P \min^2(\mathbf{o}_s(\mathbf{T}(\mathbf{p}_i, f(\mathbf{o}_i, \boldsymbol{\theta}))), 0),$$

where \mathbf{T} is the forward kinematics function of the gripper transforming \mathbf{p}_i to its global coordinates. We also assume \mathbf{o}_s has positive values outside \mathcal{O} and negative values inside

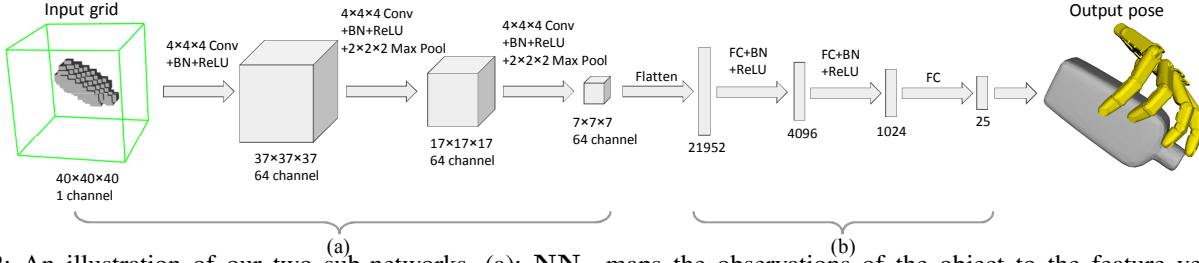


Fig. 2: An illustration of our two sub-networks. (a): NN_O maps the observations of the object to the feature vector ω . (b): NN_x maps the feature vector ω to the grasp pose x . We use the same network architecture for different robot hand hardware with different DOFs by modifying only the output layer.

\mathcal{O} . Again, $\mathcal{L}_{collision}$ is not uniformly differentiable but has a well-defined sub-gradient, so it can be used to optimize the neural network. In our experiments, we find that the quality of the learned grasp poses is sensitive to the selection of sample points p_i . We choose to use the same set of sample points for dataset generation and the collision loss function. Specifically, we use simulated annealing [9] to generate groundtruth grasp poses. [9] optimizes an approximate grasp quality function that measures the distance between a set of desired contact points to the object surfaces. These contact points are also used as sample points in $\mathcal{L}_{collision}$.

D. Combined Loss

The consistency loss and the collision loss are combined using parameters β as shown in the following equation:

$$\mathcal{L}_{combined} = \beta * \mathcal{L}_{consistency} + (1 - \beta) * \mathcal{L}_{collision}, \quad (2)$$

where the relative weight β is between 0 and 1. Empirically, we find that grasp results of higher values β are more like *GraspIt!* groundtruth, while results of lower values β bring the fingers closer to the surfaces of objects, which in turn results in higher success rates.

E. Pose Refinement

After a neural network predicts a nominal grasp pose for an unknown object, we can further refine it at runtime by looking for another grasp pose that is close to the nominal pose but does not have any intersections with the object. To do this, we solve a simple optimization. Specifically, after the neural network predicts $f(o, \theta)$, we first search for another pose x^* closest to it by minimizing the following objective function:

$$\underset{x^*}{\operatorname{argmin}} \quad \beta * \|x^* - f(o, \theta)\|^2 + (1 - \beta) * \mathcal{L}_{collision}.$$

We call this procedure pose refinement.

V. IMPLEMENTATION AND PERFORMANCE

In this section, we provide more details about our experiment platform setup, results, and evaluations.

A. Grasp Training Dataset Generation

Given a set of target objects, we take three steps to generate our grasp pose training dataset. First, we use an existing sampling-based motion planner, *GraspIt!* [26], to generate many high-quality grasp poses for each object. We then perform data augmentation via global rigid transformation. Finally, we compute a signed distance field for each of the target objects.

1) Grasp Pose Generation:

We collect object models from several datasets, including BigBIRD [33], KIT Object Models Database [15], YCB Benchmarks [6], and Grasp Database [14]. Our dataset contains $N = 324$ mesh models, of which most are everyday objects. Our high-DOF gripper

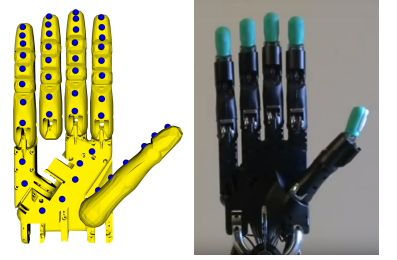


Fig. 3: Left: The Shadow Hand model we use and the sampled potential contact points in blue. Right: The real Shadow Hand.

is the Shadow Hand with 25 DOFs, as shown in Figure 3. Given an initial pose of the Shadow Hand, *GraspIt!* uses an optimization-based planner to find an optimal grasp pose that minimizes a cost function, which is found via simulated annealing. The cost function can take various forms and we use the sum of distances between sample points and object surfaces as the cost function. We run simulated annealing for 10000 iterations, where the planner generates and evaluates 10 candidate grasp poses during each iteration. To generate many redundant grasp poses for each object, we run the simulated annealing algorithm for $K = 100$ times from random initial poses. Altogether, the groundtruth grasp pose dataset is generated by calling the simulated annealing planner 324×100 times. Some grasp poses for an object are illustrated in Figure 4.

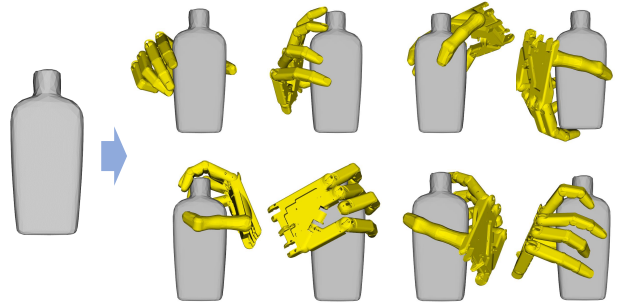


Fig. 4: An illustration of some sample grasp poses (yellow) for a single object (gray).

2) *Data Augmentation:* Generating groundtruth grasp poses using a motion planner is very computationally costly, so we use a simple method to synthesize more data. The input to NN_v is a voxelized occupancy grid. We move the

objects, put their centers of mass at the origin of the Cartesian coordinate system, and then rotate each object along with its 100 best grasp poses along 27 different rotation angles and axes. These 27 rotations are derived by concatenating rotations along X, Y and Z-axes for $60^\circ, 120^\circ, 180^\circ$, as illustrated in Figure 5. For each rotation, we record the affine transformation matrix \mathbf{T}_r . In this way, we generate a dataset that is 27 times larger than the original dataset. This data augmentation not only helps resist over-fitting when training neural networks but also helps make the neural network invariant to target object poses.

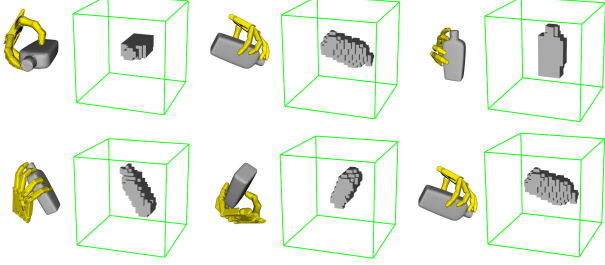


Fig. 5: An illustration of rotated object poses and grasp poses generated by data augmentation.

3) *Signed Distance Fields Construction*: To calculate the collision loss when training our neural networks, we compute a signed distance field \mathbf{G}_{sdf} for each target object by solving the Eikonal equation. We set the resolution of \mathbf{G}_{sdf} to 128^3 and \mathbf{G}_{sdf} has a local coordinate system where \mathbf{G}_{sdf} occupies the unit cube between $[0, 0, 0]$ and $[1, 1, 1]$. If the maximal length of the object's bounding box is L , the transformation matrix from an object's local coordinate system to \mathbf{G}_{sdf} 's local coordinate system is:

$$\mathbf{T}_{sdf} = \begin{bmatrix} s & 0 & 0 & 0.5 \\ 0 & s & 0 & 0.5 \\ 0 & 0 & s & 0.5 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad s \triangleq 0.95/L,$$

which is illustrated in Figure 6.

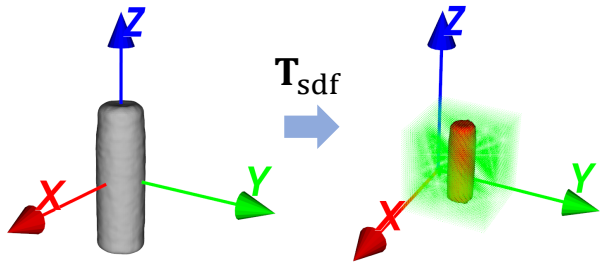


Fig. 6: Signed distance fields construction, where the green area is the domain of \mathbf{o}_s .

B. Experimental setup

We split the set of 324 target objects into an 80% (259) training set and a 20% (65) testing set. Note that each object mesh is augmented to 27 meshes with different \mathbf{T}_r . After we voxelize the meshes into 3D occupied grids, we get a total of $324 \times 27 = 8748$ grids, each with a related \mathbf{T}_r . All augmented meshes related to the same object share the same \mathbf{G}_{sdf} . Transformation from augmented meshes to \mathbf{G}_{sdf} is

given as $\mathbf{T}_{sdf} \cdot \mathbf{T}_r^{-1}$. On the Shadow Hand, we sample $P = 45$ potential contact points, as shown in Figure 3. To sample the signed distance field using $\mathbf{T}(\mathbf{p}_i, f(\mathbf{o}_i, \boldsymbol{\theta}))$, we need to transform the point from the global coordinate system to the coordinate system of the signed distance field, which is:

$$\mathbf{T}_{sdf} \cdot \mathbf{T}_r^{-1} \mathbf{T}(\mathbf{p}_i, f(\mathbf{o}_i, \boldsymbol{\theta})), \quad (3)$$

as shown in Figure 7. All experiments are carried out on a desktop with an Intel® Xeon W-2123 @ 3.60GHz \times 4, 32GB RAM, and an NVIDIA® Titan Xp graphics card with 12GB memory, on which training the neural networks takes 40 hours.

C. Results and evaluation

In this section, we evaluate the performance of our novel training method and demonstrate its benefits method for solving high-DOF grasp problems.

1) *Challenge of High-DOF Grasp Problems*: In our first benchmark, we highlight the challenges of dealing with high-DOF grippers and the necessity of our novel loss function for solving the problem. We first train our neural network using conventional supervised learning. In other words, we create a small dataset with each object corresponding to only one grasp pose ($K = 1$), and we use the simple L_2 loss function:

$$\mathcal{L}_2 = \sum_i \|f(\mathbf{o}_i, \boldsymbol{\theta}) - \mathbf{x}_i\|^2 / N.$$

With this loss function, we train two neural networks to represent grasp poses for both a high-DOF gripper (25-DOF Shadow Hand) and a low-DOF gripper (11-DOF Barrett Hand) and compare the residual of \mathcal{L}_2 after training. Due to pose ambiguity, supervised learning using the L_2 loss function can lead to inconsistency problems. Our experimental results in Table I also show that this inconsistency problem is more serious in high-DOF grippers. These two neural networks are trained using the ADAM algorithm [17] with a fixed learning rate of 0.001, a momentum of 0.9, and a batch size of 16.

Hand	DOFs of Grippers	Residual of \mathcal{L}_2 on Test Set	Residual of \mathcal{L}_2 on Training Set
Shadow	25	73.61	76.02
Barrett	11	5.84	4.76

TABLE I: We train two neural networks using an L_2 loss function to represent grasp poses for the Shadow Hand and the Barrett Hand. The residual is much higher for the Shadow Hand on both the training set and the test set, meaning that high-DOF grippers suffer more from the inconsistency problem. This can be resolved using the consistency loss function.

2) *Consistency and Collision Loss*: As shown in Table II, we train the neural network using our large dataset with $K = 100$. In this experiment, we train three neural networks using two different loss functions, $\mathcal{L}_{consistency}$ and $\mathcal{L}_{combined}$, where we pick $\beta = 0.75$. We have tried multiple choices of β and found that 0.75 leads to the best results. After training each neural network, we evaluate it on the test set and summarize the residuals of different losses, leading to 6 values in Table II; we also copy the first row of Table I to Table II as a reference for simple supervised learning method.

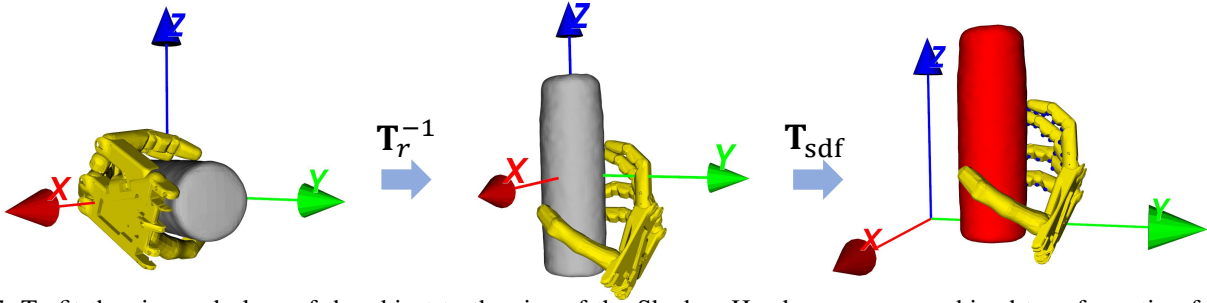


Fig. 7: To fit the size and place of the object to the size of the Shadow Hand, we use a combined transformation from the global coordinate system to \mathbf{G}_{sdf} 's local coordinate system.

Note that $\mathcal{L}_{consistency}$ and \mathcal{L}_2 both represent the distance from the neural-network-predicted grasp pose to a certain groundtruth pose, the only difference is that we have only one groundtruth pose in \mathcal{L}_2 and we have K groundtruth poses in $\mathcal{L}_{consistency}$, so $\mathcal{L}_{consistency}$ and \mathcal{L}_2 are comparable.

From the first row of Table II, we can see that, even when simple supervised learning is used at training time, the residual of $\mathcal{L}_{consistency}$ (2.630) is already much smaller than the residual of \mathcal{L}_2 (73.61). This means that the distance between the neural-network-predicted grasp pose and the closest groundtruth pose is much smaller than the average distance to all the 100 candidate grasp poses. If $\mathcal{L}_{consistency}$ is used as a loss function during training time, the residual of $\mathcal{L}_{consistency}$ is further reduced from 2.630 to 0.043. However, using combined loss does not further reduce residual metrics. In the next section, we will see that collision loss will result in grasp poses that are closer to object surfaces, which increases the success rate of grasping.

Residual	\mathcal{L}_2	$\mathcal{L}_{consistency}$	$\mathcal{L}_{combined}$ ($\beta = 0.75$)
Loss			
\mathcal{L}_2	73.61	2.630	55.865
$\mathcal{L}_{consistency}$	0.914	0.043	0.261
$\mathcal{L}_{combined}$ ($\beta = 0.75$)	0.345	0.062	0.133

TABLE II: We train neural networks using 3 different loss functions (different rows). After training, we summarize the residuals of different loss functions on the test set (different columns). Our consistency loss function drastically reduces the error of neural networks in representing a single grasp pose.

3) *Penetration Handling*: Given an object, we first use the neural network to compute a proposed grasp pose. However, this grasp pose can be invalid and may have some penetrations into the target object. We can fix this problem by combining two methods. The first method is introducing collision loss at training time. From Table II, we can see that introducing collision loss does not improve different residuals in general. However, it is very efficient in resolving most penetrations. In our experiment, introducing collision loss leads to an average relative change of the learned grasp pose by:

$$\frac{\|f_{+collision}(\mathbf{o}_i, \boldsymbol{\theta}) - f_{-collision}(\mathbf{o}_i, \boldsymbol{\theta})\|}{\|\mathbf{x}_{i,j}\|} = 12.7\%.$$

When testing the neural network trained without collision loss on the test set, an average of 2.563 of the 45 sample

points have penetrations with the target object on average and the penetration depth is $0.0553m$. With collision loss, the average number of sample points with penetration is reduced to 0.719 and the average penetration depth is reduced to $0.0081m$. However, there are still some small penetrations, as shown in Figure 9 (a). During runtime, the actual grasping hardware cannot allow any penetrations between the object and the Shadow Hand. A second method is needed to compute a grasp pose without any penetrations; we use a simple interpolation method (runtime adjustment). Specifically, we compute the gradient of $\mathcal{L}_{collision}$ with respect to the joint angles:

$$\frac{\partial [\sum_{i=1}^P \min^2(\mathbf{o}_s(\mathbf{T}(\mathbf{p}_i, \mathbf{x})), 0)]}{\partial \mathbf{x}}$$

and we update our joint pose along the negative gradient direction until there are no penetrations. In practice, a single forward propagation through the neural network takes a computational time of $0.541s$ and the runtime adjustment takes $0.468s$.

4) *Multi-View Depth Image as Input*: To extend our method to real-world scenarios, we evaluate our method on object models with uncertainties or inaccuracies. The real Shadow Hand is mounted on a UR10 arm and we use Shadow Robot Interface to move the arm and hand. In our experiment, we select 15 objects from the YCB Benchmarks, none of which have been included in our training dataset or test dataset. These 15 objects are captured using a multi-view depth camera and their geometric shapes are constructed using the standard pipeline implemented in [31] and illustrated in Figure 8. Specifically, RANSAC is first used to remove the planar background of the obtained point cloud, Euclidean cluster extraction algorithm is used to find a set of segmented object point clouds, and segmented object meshes are then extracted using Poisson surface reconstruction. The reconstructed meshes are finally voxelized to a 3D occupancy grid. After pre-process, reconstructed mesh is fed to our neural networks to generate grasp poses. Sometimes the generated grasp poses are of low quality in terms of the ϵ -metric, in which case we rotate the object mesh and run our neural networks again to generate a new grasp pose. On average, we rotate the object 3-5 times and report the best grasp pose quality in the wrench space. Although these reconstructed object meshes have noisy surfaces, we still get an average grasp quality of 0.102 over the 15 objects where we use the ϵ -metric to measure grasp quality. Some grasp

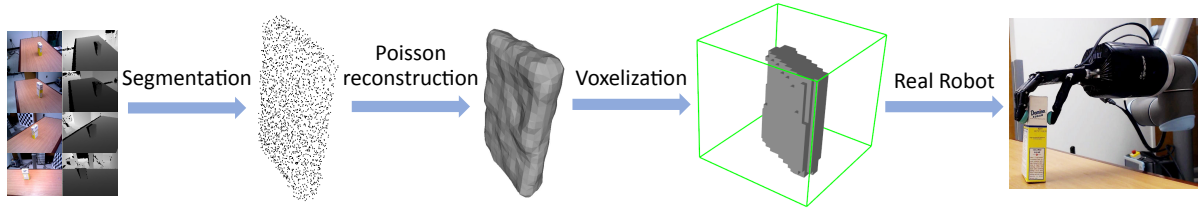


Fig. 8: Our neural network can use inaccurate object models reconstructed from multi-view depth images. The object meshes are reconstructed by first segmenting the point cloud and excluding the background, then applying Poisson surface reconstruction, and finally voxelizing the model.

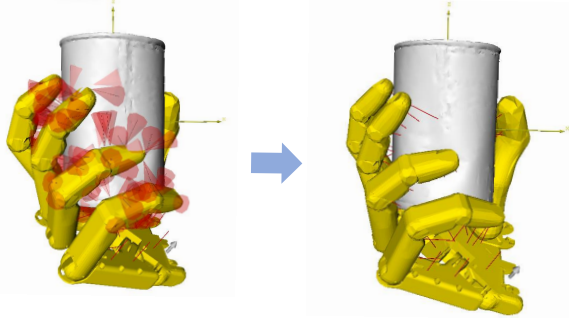


Fig. 9: There are still some penetrations after $\mathcal{L}_{collision}$ is used (a), and we can resolve these penetrations (b) by using pose refinement during runtime.

poses are shown in Figure 10. We carry out our real robot grasping for 15 objects after finding a feasible grasp for each and 12 are successful.

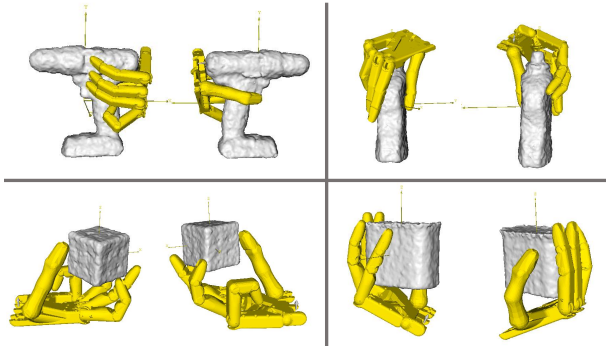


Fig. 10: High-quality grasp poses from 2 different views for 4 objects.

5) *Comparison with Prior Methods:* The main difference between our method and prior works [10], [36], [7] is that we target high-DOF grasp poses and we use a neural network to generate grasp poses directly instead of using the score of a candidate grasp pose. Our method still needs a sampling-based algorithm to randomly rotate the target object and pick the best grasp. However, unlike [10], which requires hundreds of samples, our method only needs 3-5 samples, which can be computed within 3-5 seconds. On the other hand, a major drawback of our method is that we require a very large dataset, with tens of grasp poses for each target object. We find this dataset to be an essential component of making our method robust when generating grasp poses for unseen objects, as shown in Figure 12. Most of the generated grasp poses (after pose refinement) for unseen objects are of qualities similar to poses generated from *GraspIt!*, as shown

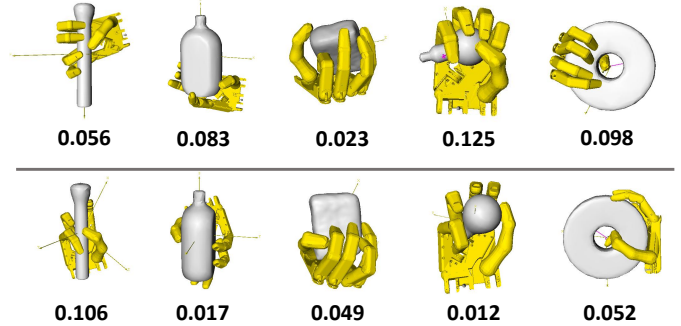


Fig. 11: A comparison of grasp pose quality generated using *GraspIt!* (top row) and our method (bottom row).

in Figure 11.

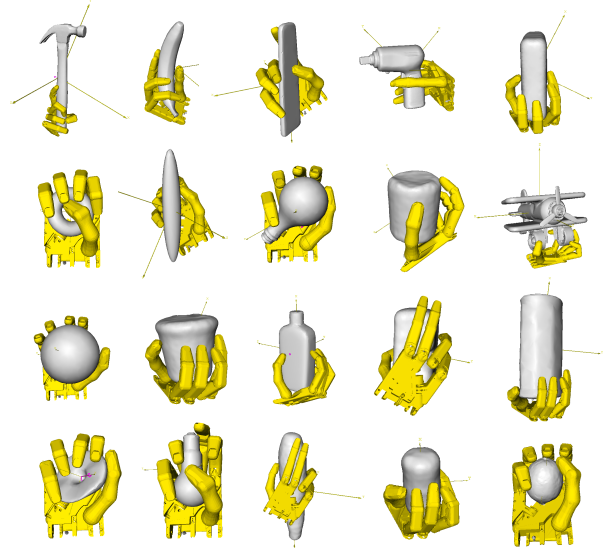


Fig. 12: Stable grasp poses generated for a large collection of unknown objects.

VI. CONCLUSION AND LIMITATIONS

We present a new neural-network architecture and a training technique for the generation of high-DOF grasp poses. To resolve grasp pose redundancy, we use a consistency loss function and let the neural network pick the best or most-representable grasp poses for each target object. To further improve the quality of the grasp poses, we introduce a collision loss function to resolve penetrations between the hand and the object. Our results show that conventional supervised learning will not result in accurate grasp poses

while a neural network trained using our consistency loss function drastically improves the accuracy of grasp poses compared to conventional supervised learning. Further, the collision loss can effectively resolve penetrations between the gripper and the target object on both the training set and the test set.

A major limitation of our current method is that it requires a very large dataset with many effective grasp poses for each target object. This is essential for the neural network to select consistent grasp poses. However, when we have a very large set of target objects, generating such a dataset can be very computationally costly and lots of computations can be wasted because the computed grasp poses are not selected by the neural network. Another limitation is that our one-to-one mapping method can cause a reachability problem. When a grasp pose is not reachable or collision-free, we have to rotate the target object grid and then feed it to our neural network until we find a feasible grasp pose.

There are several avenues for future work. One is to consider an end-to-end architecture that predicts grasp poses directly from multi-view depth images, similar to [36]. Another direction is to consider more topologically complex target objects, such as high-genus models. In these cases, a signed distance representation is not enough to resolve the geometric details of objects and the collision loss needs to be reformulated. Finally, our current method has been evaluated on a single high-DOF gripper model (the Shadow Hand) and it would be useful to generalize the ability of the neural network to represent grasp poses for other high-DOF gripper models such as a humanoid hand model, in which case we could utilize a prior method [3] to generate humanoid grasp data.

VII. ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their valuable comments. This work was supported in part by NSFC (61572507, 61532003, 61622212) and Intel. Min Liu is supported by the China Scholarship Council.

REFERENCES

- [1] *Grasp Evaluation with Graspable Feature Matching*, 2011.
- [2] M. Andrychowicz, B. Baker, M. Chociej, R. Józefowicz, B. McGrew, J. W. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, J. Schneider, S. Sidor, J. Tobin, P. Welinder, L. Weng, and W. Zaremba, "Learning dexterous in-hand manipulation," *CoRR*, vol. abs/1808.00177, 2018.
- [3] A. Betancourt, P. Morerio, L. Marcenaro, E. Barakova, M. Rauterberg, and C. Regazzoni, "Towards a unified framework for hand-based methods in first person vision," in *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2015, pp. 1–6.
- [4] C. Borst, M. Fischer, and G. Hirzinger, "A fast and robust grasp planner for arbitrary 3d objects," in *Proceedings 1999 IEEE International Conference on Robotics and Automation (Cat. No.99CH36288C)*, vol. 3, May 1999, pp. 1890–1896 vol.3.
- [5] K. Bousmalis, A. Irpan, P. Wohlhart, Y. Bai, M. Kelcey, M. Kalakrishnan, L. Downs, J. Ibarz, P. Pastor, K. Konolige, S. Levine, and V. Vanhoucke, "Using simulation and domain adaptation to improve efficiency of deep robotic grasping," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018, pp. 4243–4250.
- [6] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Benchmarking in manipulation research: The ycb object and model set and benchmarking protocols," *IEEE Robotics and Automation Magazine*, pp. 36–52, 2015.
- [7] C. Choi, W. Schwarting, J. DelPreto, and D. Rus, "Learning object grasping for soft robot hands," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2370–2377, July 2018.
- [8] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, "3d-r2n2: A unified approach for single and multi-view 3d object reconstruction," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.
- [9] M. Ciocarlie, C. Goldfeder, and P. Allen, "Dimensionality reduction for hand-independent dexterous robotic grasping," in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2007, pp. 3270–3275.
- [10] K. Fang, Y. Bai, S. Hinterstoisser, S. Savarese, and M. Kalakrishnan, "Multi-task domain adaptation for deep learning of instance grasping from simulation," *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3516–3523, 2018.
- [11] R. Girdhar, D. Fouhey, M. Rodriguez, and A. Gupta, "Learning a predictable and generative vector representation for objects," in *ECCV*, 2016.
- [12] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from rgbd images: Learning using a new rectangle representation," *2011 IEEE International Conference on Robotics and Automation*, pp. 3304–3311, 2011.
- [13] E. Johns, S. Leutenegger, and A. J. Davison, "Deep learning a grasp function for grasping under gripper pose uncertainty," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2016, pp. 4461–4468.
- [14] D. Kappler, B. Bohg, and S. Schaal, "Leveraging big data for grasp planning," in *Proceedings of the IEEE International Conference on Robotics and Automation*, May 2015.
- [15] A. Kasper, Z. Xue, and R. Dillmann, "The kit object models database: An object model database for object recognition, localization and manipulation in service robotics," *The International Journal of Robotics Research*, vol. 31, no. 8, pp. 927–934, 2012.
- [16] N. Ketkar, "Introduction to pytorch," in *Deep Learning with Python*. Springer, 2017, pp. 195–208.
- [17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [18] M. Kopicik, R. Detry, M. Adjigbe, R. Stolkin, A. Leonardis, and J. L. Wyatt, "One-shot learning and generation of dexterous grasps for novel objects," *The International Journal of Robotics Research*, vol. 35, no. 8, pp. 959–976, 2016.
- [19] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *The International Journal of Robotics Research*, vol. 34, no. 4–5, pp. 705–724, 2015.
- [20] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *The International Journal of Robotics Research*, vol. 37, no. 4–5, pp. 421–436, 2018.
- [21] Q. Lu, K. Chenna, B. Sundaralingam, and T. Hermans, "Planning multi-fingered grasps as probabilistic inference in a learned deep network," in *Intl Symp. on Robotics Research*, 2017.
- [22] Q. Lu and T. Hermans, "Modeling Grasp Type Improves Learning-Based Grasp Planning," *IEEE Robotics and Automation Letters*, 2019.
- [23] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, and K. Goldberg, "Learning ambidextrous robot grasping policies," *Science Robotics*, vol. 4, no. 26, p. eaau4984, 2019.
- [24] J. Mahler, F. T. Pokorny, B. Hou, M. Roderick, M. Laskey, M. Aubry, K. Kohlhoff, T. Kröger, J. Kuffner, and K. Goldberg, "Dex-net 1.0: A cloud-based network of 3d objects for robust grasp planning using a multi-armed bandit model with correlated rewards," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 1957–1964.
- [25] A. T. Miller and P. K. Allen, "Examples of 3d grasp quality computations," in *Proceedings 1999 IEEE International Conference on Robotics and Automation (Cat. No.99CH36288C)*, vol. 2, May 1999, pp. 1240–1246 vol.2.
- [26] A. Miller and P. Allen, "Graspit! a versatile simulator for robotic grasping," *Robotics Automation Magazine, IEEE*, vol. 11, no. 4, pp. 110–122, Dec. 2004.
- [27] L. Montesano and M. Lopes, "Active learning of visual descriptors for grasping using non-parametric smoothed beta distributions," *Robot. Auton. Syst.*, vol. 60, no. 3, pp. 452–462, Mar. 2012.
- [28] S. Osher and R. Fedkiw, *Level set methods and dynamic implicit surfaces*. Springer Science & Business Media, 2006, vol. 153.
- [29] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, May 2016, pp. 3406–3413.
- [30] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine, "Learning Complex Dexterous Manipulation with Deep Reinforcement Learning and Demonstrations," in *Proceedings of Robotics: Science and Systems (RSS)*, 2018.
- [31] R. B. Rusu and S. Cousins, "3d is here: Point cloud library (pcl)," in *2011 IEEE International Conference on Robotics and Automation*, May 2011, pp. 1–4.
- [32] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *The International Journal of Robotics Research*, vol. 27, no. 2, pp. 157–173, 2008.
- [33] A. Singh, J. Sha, K. S. Narayan, T. Achim, and P. Abbeel, "Bigbird: A large-scale 3d database of object instances," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 509–516.
- [34] N. Vahrenkamp, S. Wieland, P. Azad, D. Gonzalez, T. Asfour, and R. Dillmann, "Visual servoing for humanoid grasping and manipulation tasks," in *Humanoids 2008 - 8th IEEE-RAS International Conference on Humanoid Robots*, Dec 2008, pp. 406–412.
- [35] Z. Xue, A. Kasper, J. M. Zöllner, and R. Dillmann, "An automatic grasp planning system for service robots," *2009 International Conference on Advanced Robotics*, pp. 1–6, 2009.
- [36] X. Yan, J. Hsu, M. Khansari, Y. Bai, A. Pathak, A. Gupta, J. Davidson, and H. Lee, "Learning 6-dof grasping interaction via deep geometry-aware 3d representations," *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–9, 2018.
- [37] A. Zeng, S. Song, K. Yu, E. Donlon, F. R. Hogan, M. Bauzá, D. Ma, O. Taylor, M. Liu, E. Romo, N. Fazeli, F. Alet, N. C. Daffe, R. Holladay, I. Morona, P. Q. Nair, D. Green, I. Taylor, W. Liu, T. A. Funkhouser, and A. Rodriguez, "Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching," in *2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21–25, 2018*, 2018, pp. 1–8.
- [38] Y. Zheng, "An efficient algorithm for a grasp quality measure," *IEEE Transactions on Robotics*, vol. 29, no. 2, pp. 579–585, April 2013.